OPENSYSTEMONE: EFFICIENT PARALLEL INFERENCE FOR DYNAMIC STRUCTURED DECISION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine-facing AI workloads often require selecting, scoring, or verifying a bounded set of alternatives rather than generating natural language. Yet such workloads are commonly served by autoregressive decoders, and whether their computational structure matches the task is rarely measured at the level of the decision. We present OpenSystemOne, an open benchmarking and systems framework for parallel structured decision inference. The framework implements and compares free and constrained autoregressive generation, option likelihood scoring with and without shared-prefix reuse, encoder-based classifiers, query-conditioned decision heads, and learned dynamic-option models under a common System One interface. We measure accuracy and calibration on four public decision benchmarks and probe reordered, reduced, gold-absent and out-of-scope candidate sets in intent routing. On one Blackwell GPU, we measure latency, throughput, device memory and GPU energy on fixed request cohorts and vary context length, option cardinality and batch size one factor at a time. Eliminating answer generation does not by itself lower cost: a one-token constrained decoder already decides after one prefill, and on fixed batch-32 cohorts it completes more decisions per second than our encoder query head (342 versus 224 on BANKING77). The largest measured change was host-side: moving a per-request tokenization check out of timed decoder calls raises throughput 13-fold for requests prepared in advance, but counting that setup leaves no demonstrated first-use gain. In FP32, the precision in which reused and fresh paths agree within tolerance, prefix reuse raises multi-question throughput at higher memory cost. Quality and calibration vary by predictor and workload, and changing candidate sets expose order sensitivity and constant abstention. Characterizing when eliminating token-by-token answer generation pays off end to end, we find that for the tested implementations cost resides in the complete execution path rather than in generated tokens, while quality on the served candidate sets decides which fast path is usable.

1 INTRODUCTION

An intent router chooses among service actions, a multiple-choice reader scores supplied answers and a verifier assigns one of a few outcomes. In each case the alternatives are fixed before inference, and the caller is software that needs a selected identifier, sometimes with a probability for every candidate, instead of free text. We call a model that answers such requests directly with typed decisions a System One model: a fast decision layer that software can invoke repeatedly and act on, as distinct from open-ended generation. A language model can produce that decision by generating a label, by scoring each candidate as a continuation or by selecting one token from a constrained vocabulary. An encoder can score a fixed label inventory or condition a decision head on candidate descriptions and generate nothing at all. These formulations differ in how often the context is processed, how much of the work runs in parallel and what the returned scores mean.

The systems question is whether such decisions need token-by-token generation at all. Parallel formulations promise the same decision for less latency and compute, but the number of generated tokens does not settle that promise. A constrained decoder whose identifiers are single tokens already selects an answer directly after prefill. A likelihood scorer evaluates candidate suffixes in parallel and can reuse their shared prefix. A query head generates nothing, yet it still encodes the state, the question and every candidate. Relative cost therefore depends on context length, candidate

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

cardinality, batch size, cache state, host-side preprocessing and whether the caller needs a full candidate distribution. Testing the promise requires comparing formulations at the level of the decision, on the same requests, candidate sets and output contract, and counting candidate encoding, host overhead and setup wherever the caller pays for them.

We organize the study around three research questions. RQ1 asks how the families differ in accuracy and calibration on retained labeled evaluation examples, and how decisions and calibration shift when candidate sets are reordered, reduced, stripped of the correct label or drawn from labels unseen in training. RQ2 asks what each family costs in latency, throughput, device memory and GPU energy on fixed inputs executed on one Blackwell GPU. RQ3 asks how much of the cost of repeated questions about one state is removed by batching and how much by prefix reuse, and what a profiled host-side overhead contributes to decoder cost.

The answers are conditional, and the conditions carry the main lesson. Quality depends jointly on predictor, adaptation and workload; the fixed classifier is more accurate than the adapted decoder on known intents and less accurate on out-of-scope queries. On fixed batch-32 cohorts, the tested one-token constrained decoder completes more decisions per second than our encoder query head on both measured tasks, so avoiding answer-token generation did not by itself make a formulation cheaper here. A large observed cost change came from host-side work, namely moving a per-request token-boundary check out of timed decoder calls, and it helps only requests certified in advance. Quality is estimated on the retained evaluation sets and cost on fixed request cohorts, and the two remain separate estimands; model scale is fixed within each measured family, and one GPU is measured. The controls are chosen to narrow attribution: free and constrained generation share an adapted model, cached and uncached scoring share weights and scoring rule, the learned heads are ablated for candidate interaction and token-level state access, and four shared-state paths change only batching and reuse. The contribution is this controlled empirical comparison and the open interface on which it runs, which lets System One models be measured against generation-based paths on identical requests; the individual components predate it. Specifically, the study provides:

- a common decision interface covering free and constrained generation, likelihood scoring with and without prefix reuse, encoder classifiers, query-conditioned heads and candidate-interacting dynamic-option models, validated by 49 CPU interface checks (Section 3);
- retained-evaluation-set accuracy and calibration for 167 primary endpoints on four public benchmarks, and 14 changing-candidate collections separating order, subset, gold-absent and out-of-scope conditions (Sections 4.1 and 4.2);
- 88 resource cells on one Blackwell GPU with device telemetry and failure accounting, including a 2×2 separation of batching from prefix reuse and one profile-guided before/after change whose setup cost is counted (Sections 4.3–4.5).

2 Related work

Single-token decisions, option scoring and serving. Robinson et al. (2023) contrast cloze prompting, which scores each candidate separately and typically needs length or unconditional normalization, with multiple-choice prompting, which selects the most probable symbol token in one pass. Our constrained one-token and option-likelihood baselines implement these formulations. SGLang supports choice selection and prefix reuse in structured language-model programs (Zheng et al., 2024), FlashInfer provides attention kernels with shared-prefix computation (Ye et al., 2025), and vLLM manages KV memory in pages (Kwon et al., 2023); vLLM serves our decoder measurements.

Candidate-conditioned classifiers. Yin et al. (2019) cast zero-shot classification as entailment between text and a label hypothesis. GLiClass jointly represents text and supplied labels (Stepanov et al., 2025), and GLiNER2 composes schema-prompted classification and extraction in one encoder pass (Zaratiana et al., 2025). SCX Router scores inference-time labels without generation by pairing a Qwen3 decoder with a shallow bidirectional scorer over a reusable text-only KV cache (Stepanov et al., 2026). We discuss GLiNER2 and SCX Router as related designs and do not evaluate them. ModernBERT is our encoder backbone (Warner et al., 2024), and our set head follows the permutation-equivariant attention of Set Transformer (Lee et al., 2019).

Inference energy and calibration. TokenPowerBench sweeps batch size, context, quantization and parallelism for generative inference with GPU, CPU/DRAM and node telemetry (Niu et al., 2026). Luccioni et al. (2024) report lower inference energy for task-specific discriminative models than for multi-purpose generative ones on classification and question answering. Our energy boundary is cumulative GPU device counters only. Fine-grained intent routing and out-of-scope detection (Casanueva et al., 2020; Larson et al., 2019) motivate separate known-label and abstention evaluation, and we pair accuracy with proper scores and held-out temperature scaling (Guo et al., 2017).

3 Decision interface and experimental overview

This section summarizes what the results require; Appendix A gives the complete protocol, including equations, prompts, stopping rules and filters.

Decision contract. A request $r = (x, q, \mathcal{C})$ holds state text x, question q and an ordered collection $\mathcal{C}$ of K candidates with stable identifiers and descriptions. Tasks that permit rejection include an explicit abstention candidate. The decision-only contract returns a selected identifier and a validity status. The full-distribution contract also returns candidate probabilities $p_i(T) \propto \exp(s_i/T)$ from real-valued scores s_i, with $T = 1$ for native probabilities and T fitted on disjoint calibration data otherwise. Invalid or truncated generations remain failed decisions, and a method that needs an extra call to supply a distribution pays for that call.

Seven families. Free generation decodes the chosen identifier greedily with at most 32 new tokens and returns decision-only output. Constrained decoding restricts transitions to identifier strings; when every identifier is verified to be a single token at the prompt boundary, a one-token path reads the allowed-token logits from one prefill. Option likelihood scoring treats each description as a continuation of the state–question prefix, using the raw or length-normalized log-probability sum chosen on tuning data. Cached likelihood scoring computes the prefix once and scores suffixes from its KV state with the same weights and rule. Encoder baselines comprise a fixed classifier over the known label inventory and a cross-encoder that scores each state–question–candidate triple. The query head encodes the state once into token representations, and all candidate queries attend to that memory in one batched call. The learned dynamic-option models are our set head, which adds self-attention across candidates without index embeddings, and released GLiClass-modern (Stepanov et al., 2025). A pooled-query control replaces the state-token memory with its mean. Our heads instantiate a public typed-decision interface, which is the sense in which we call them Jev-like (TypeSafe AI, n.d.); they do not reproduce proprietary Jev weights, architecture or training.

Token-boundary certification and runtimes. Decoder requests must be tokenized so that the answer boundary preserves the intended prefix tokens; we call this check token-boundary certification. In the measured decoder paths it runs once per request at startup, outside timed calls, and its time is reported separately. Certified mode rejects a changed or unseen request until it is prepared. Decoder resource paths run in vLLM in BF16. Encoder paths run in our PyTorch implementation with BF16 autocast and tokenize inside timed calls.

Workloads and endpoints. The suite comprises BANKING77 and CLINC150 intent routing, with CLINC150 out-of-scope (OOS) queries as a rejection stratum (Casanueva et al., 2020; Larson et al., 2019), SciQ answer selection (Welbl et al., 2017) and BoolQ passage-conditioned verification (Clark et al., 2019). Leakage groups keep training, tuning, calibration and test disjoint (Table 3). SciQ questions with originally missing support form a separate stratum, and a stable hash partition of semantic labels yields held-out-label test views of BANKING77 (13 labels) and CLINC150 (34 labels). Released endpoints include Qwen3-0.6B (Yang et al., 2025) and GLiClass-modern. Adapted endpoints use Qwen3-0.6B with LoRA and fully adapted ModernBERT-base (Warner et al., 2024) under fixed budgets with tuning-only selection. Twenty-six primary families are trained at seeds 17, 29 and 43, and supporting configurations are single-seed. The result is 167 primary canonical endpoints at batch size 1 and 22 scoring-rule sensitivity endpoints; coverage follows where each family applies and is not a full method-by-dataset product.

Quality and changing candidates. Accuracy and macro F1 keep invalid outputs in the denominator. NLL and Brier score are the primary probability metrics, and a scalar temperature (Guo et al., 2017) is fitted by NLL on the calibration split only. Descriptive 95% intervals use 2,000 cluster-bootstrap resamples over test groups; seeds are reported separately and no multiplicity-adjusted superiority claim is made. Changing-candidate probes run seven methods at seed 17 on 128 BANKING77 held-out-label groups and on 96 in-scope and 32 OOS CLINC150 groups: 14 jobs and 13,664 requests at batch size 32. Conditions are one seeded reordering, gold-included and gold-absent subsets of size K, and OOS subsets. Subsets contain the abstention candidate `__abstain__`, which the canonical BANKING77 set lacks, and gold-absent accuracy is the abstention rate. Each endpoint's canonical temperature is reused unchanged.

Resource measurements. All 88 cells run on one of four NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs with fixed CPU affinity. Natural-input cells (42) measure 21 endpoint identities at batch sizes 1 and 32 on two 32-request cohorts fixed by request identifier before collection: BANKING77 known-label in-scope and SciQ passage-conditioned requests. One-factor cells (30) vary state tokens, K or batch size around an anchor of $(512, 8, 1)$ with constructed requests. Shared-state cells (16) cross fresh versus reused-prefix and serial versus batched execution for released Qwen3-0.6B in FP32, with $Q \in \{1, 16\}$ questions on two fixture states. Each cell has five blocks with an excluded warmup and a matched idle measurement. Timing covers tokenization, transfer, inference, normalization and output assembly, and excludes loading, compilation, warmup and startup certification. Energy is the difference of cumulative GPU counters over active windows, reported gross and idle-adjusted per decision. Blocks with a foreign process on the measured GPU or missing telemetry are excluded without replacement, and a direction is stated only when the difference of block medians exceeds the larger within-visit block range. Decoder visits were collected again in a second session after the optimization below. A bridge cell measured in both sessions suppresses cross-session directions within its shift or in metrics it lacks, including every batch-1 request-latency direction.

Profiling and one optimization. Python cProfile ranked host components in the BANKING77 batch-32 constrained and query-head cells by exclusive wall-clock self time; it does not observe GPU kernels. The selected component was per-request certification inside every timed decoder call. A matched before/after trial on one GPU, cell and request set moved it to startup and compared first-observation selections, statuses and scores. The change was then applied to all decoder paths, and encoder paths were left unchanged.

4 RESULTS

4.1 INTERFACE, QUALITY AND CALIBRATION

All families run behind the common contract. Free generation returns decision-only output and every other family returns full distributions. The CPU interface suite comprises 49 checks across nine modules, all of which are satisfied by the evaluated implementation. These checks cover implementation contracts such as candidate identity and scoring boundaries; they neither show that predictors compute equivalent scores nor reproduces the GPU runs, and several BF16 GPU comparisons fail (Sections 4.4 and 4.5).

Quality depends on predictor, adaptation and workload together (Table 1; all rows in Table 4). On BANKING77 the fixed classifier (0.905) and adapted likelihood scoring (0.864) have higher point accuracy than the constrained decoder (0.804), the query head (0.733) and the set head (0.648). Released decoders without adaptation are weak on the intent tasks and some fail degenerately; released likelihood scoring abstains on nearly every CLINC150 request. Larger released Qwen3 checkpoints raise constrained BANKING77 accuracy from 0.203 (0.6B) to 0.379 (1.7B) and 0.556 (4B), all below the adapted 0.6B decoder.

We also paired the adapted constrained decoder (Qwen3-0.6B with LoRA in vLLM) with fully fine-tuned ModernBERT-base encoders on identical requests (Appendix C). These six comparisons were chosen after per-side accuracies were visible, and their intervals are unadjusted and descriptive. At each of the three trained seeds, the fixed classifier is more accurate on known intents, by 9.4–11.3 percentage points on BANKING77 and 6.7–9.8 on CLINC150 in scope, whereas the decoder is more

	BANKING77	CLINC150		SciQ		BoolQ
Endpoint		in scope	OOS	passage	missing	
n	3,079	4,498	1,000	879	111	3,270
Adapted constrained	0.804	0.886	0.909	0.994	0.793	0.812
Fixed classifier	0.905	0.955	0.730	–	–	–
Cross-encoder	0.888	–	–	0.994	0.559	0.761
Query head	0.733	0.862	0.718	0.474	0.550	0.658
Set head	0.648	0.749	0.844	0.667	0.468	0.642

Table 1: Accuracy on the retained evaluation examples (BoolQ: official validation split) at seed 17 in the known-label regimes. The constrained decoder is Qwen3-0.6B with LoRA; the encoder paths are fully adapted ModernBERT-base. SciQ strata are passage-conditioned and missing-support questions. Invalid and truncated outputs count as errors, and a dash marks a configuration not evaluated. Table 4 adds generation, likelihood and released endpoints.

accurate on CLINC150 OOS queries, by 13.1–20.7 points. At seed 17 the cross-encoder trails the decoder by 5.1 points on BoolQ and ties it at 0.994 on SciQ passage-conditioned questions. This tie sits at a ceiling and does not establish equivalence. Wherever accuracy differs, native NLL and Brier score favor the same system. The compared systems differ in backbone, adaptation and runtime, and a fixed classifier cannot score labels outside its inventory. These batch-1 quality differences cannot be combined with the cohort costs of Section 4.3 into a quality–cost frontier.

Replicate stability differs by method. The BANKING77 fixed classifier gives 0.905, 0.922 and 0.912 at seeds 17, 29 and 43, whereas the CLINC150 query head gives 0.862, 0.768 and 0.683 in scope and 0.718, 0.267 and 0.202 on OOS queries. The SciQ query head stays near 0.47 on passage-conditioned questions at all three seeds (0.474, 0.478, 0.481), while the seed-17 cross-encoder reaches 0.994 on that stratum. On naturally missing-support questions the query head scores 0.550, 0.414 and 0.495. Seed 17 is thus lower with a passage, seed 29 reverses the direction and seed 43 is nearly level. The strata contain different questions, so this comparison is not a paired passage-removal experiment; we did not measure passage use and draw no conclusion that passages harm the heads.

Held-out temperature scaling does not uniformly improve NLL. It helps released constrained decoding on BANKING77 (8.66 to 3.89) but worsens NLL for the seed-17 set head on CLINC150 held-out labels (2.915 to 3.105) and on some CLINC150 OOS strata, for example the seed-29 query head (2.389 to 2.652) and released GLiClass-modern (4.881 to 5.407). We computed no paired interval for calibration effects, so these are observed differences without a significance claim. The CLINC150 calibration split contains 56 OOS examples among 1,538, against 1,000 among 5,498 in test. This prevalence difference limits interpretation of OOS calibration transfer, and the design does not identify its contribution. On the held-out-label views, adapted constrained decoding reaches 0.860 (BANKING77) and 0.901 (CLINC150) and released GLiClass-modern 0.608 and 0.542, whereas the query head reaches 0.290 and 0.121 and the set head 0.062 and 0.204; these values refer only to our label partition. The adapted decoders select among supplied identifiers after training on other labels, whereas the learned heads, under our budgets, transfer poorly to this partition.

4.2 Changing candidate sets

Reordering candidates changes decisions in some predictors and not others (Table 5). Under the tested permutation, the query and set heads change no selected identifier, consistent with their equivariant construction, and the likelihood scorers change at most two. The adapted decoders change 9 of 128 BANKING77 requests and 6–7 of 96 CLINC150 in-scope requests, and GLiClass-modern changes 26 of 128, 43 of 96 and 18 of 32 OOS requests. Paired intervals for the accuracy change include zero in these cases, for example -0.016 $[-0.078, 0.047]$ for GLiClass-modern on BANKING77. One reordering thus alters individual decisions without a detected shift in average accuracy; this is no evidence about invariance across orderings or seeds.

Some high gold-absent accuracies come from constant outputs. The adapted uncached and cached likelihood scorers select abstention for every BANKING77 request in every tested subset, scoring

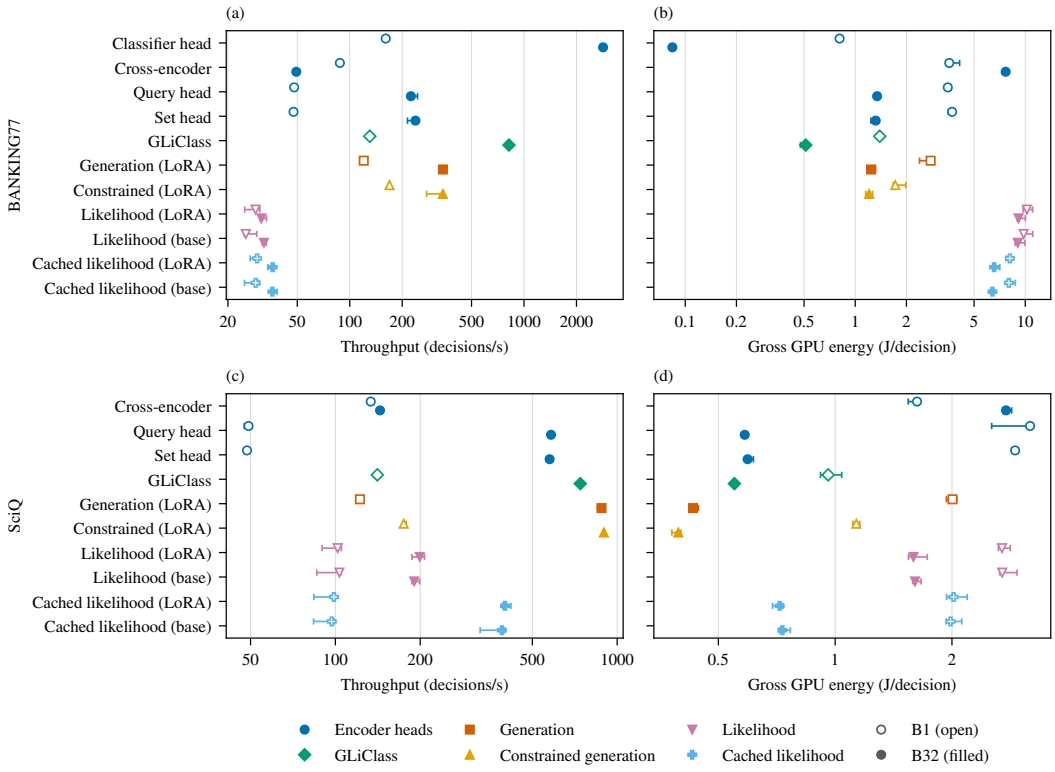

Figure 1: Throughput (decisions/s) and gross GPU energy (J/decision) for the 42 natural-input cells on the fixed 32-request BANKING77 and SciQ cohorts, measured on one Blackwell GPU. Rows are endpoint identities; hollow markers show batch size 1 and filled markers batch size 32. Markers are medians over five blocks and bars are observed block ranges (minimum to maximum), not confidence intervals. Decoder cells come from the repaired session, with per-request token-boundary certification run at startup and excluded from measurement; encoder cells come from the original session. The two BANKING77 cells of released uncached likelihood scoring come from a visit classified as incomplete execution.

1.000 when the gold label is absent and 0.000 when it is included. The set head does the same at $K = 8$ and $K = 13$. These are failures to discriminate labels once abstention is offered, not evidence of rejection ability or label transfer. Elsewhere abstention varies with cardinality: GLiClass-modern abstains on 0.740, 0.313 and 0.208 of CLINC150 gold-absent requests at $K = 2$, 8 and 32, while the adapted decoders abstain on 0.948 at $K = 32$. Subset contrasts also change batch companions and are not attributed to candidates alone. Applying the canonical temperature raises NLL in some conditions, for example from 1.593 to 1.787 for GLiClass-modern on CLINC150 gold-absent $K = 8$; without paired calibration intervals, these results establish neither calibration transfer nor its general failure.

4.3 Serving costs on fixed cohorts

Figure 1 shows the 42 natural cells, and Table 6 lists batch-32 throughput beside full-test accuracy. Across the plotted batch sizes, medians span more than an order of magnitude within each cohort. Our encoder-versus-decoder conclusion rests on two registered contrasts, adapted constrained decoding against the fully adapted query head at batch size 32. On BANKING77 the constrained decoder completes 342.3 decisions/s and the query head 224.4; the 117.9 decisions/s difference exceeds the larger within-visit block range of 71.2, set by one slow decoder block. On SciQ the medians are 897.3 and 582.3 decisions/s with non-overlapping block ranges. On these fixed, pre-certified cohorts, the tested one-token decoder therefore has higher observed throughput than the tested en-

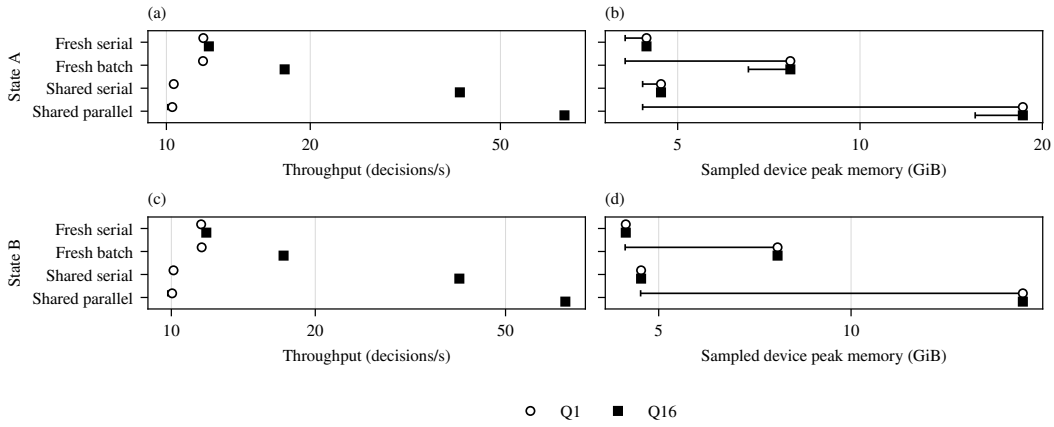

Figure 2: Throughput (decisions/s) and sampled device memory (GiB) for the 16 FP32 shared-state cells with released Qwen3-0.6B. The four paths are fresh serial, fresh batched, reused-prefix serial and reused-prefix parallel. States A and B are the shortest and median fixture states; circles mark $Q = 1$ and squares $Q = 16$ questions per call, and each call returns Q decisions. Markers are medians over five blocks and bars are observed block ranges, not confidence intervals. Every path repeats the same per-question prompt certification inside timed calls. The fixture is unlabeled, so the cells carry no quality information.

coder head. Batch-1 call and request latencies are descriptive, and no cross-session time-to-decision ordering is drawn from them.

This is a comparison of tested implementations. The decoder runs in vLLM and the query head in PyTorch, only decoder paths received the preprocessing change of Section 4.5, so the query head still tokenizes inside timed calls, and the backbones differ in pretraining and adaptation. Decoder rates exclude certification, about 37 ms per request in the representative BANKING77 trial. The two sides of each contrast were collected about 24,000 s apart. The bridge cell's throughput shifted by 0.0059 decisions/s against its own 0.334 decisions/s block range, but it measures the unrepaired path near 25.8 decisions/s and carries little drift information at 200–900 decisions/s. Directions that survive this screen are merely not suppressed by it; they are not shown to be robust to drift.

The repaired collections changed some decisions relative to the original session: selected identifiers differ in 3 of 42 repaired cells, with no status disagreements. The BANKING77 batch-32 constrained cell is one of them. One of 32 requests differs in each of two of its five block pairs, and the two pairs involve different requests. One of these requests also changes selection across repeated calls on both sides of the optimization trial; the other changes across repeated calls within repaired block index 4. We therefore make no quality-preservation statement for the batch-32 cohorts, and the retained-evaluation-set accuracies (0.804 versus 0.733 on BANKING77, 0.994 versus 0.474 on SciQ) are a separate batch-1 estimand.

In the BANKING77 contrast the idle-adjusted GPU energy difference of 0.011 J per decision lies below its 0.116 J floor, so no energy ordering follows for that contrast. The vLLM paths report zero parent-process allocator peaks, which we treat as unavailable, and their sampled device occupancy (about 78.8 GB in the BANKING77 constrained cell) includes the engine's preallocated pool. It cannot be compared as model memory with the 2.1–7.4 GB sampled for encoder paths. The visit of released uncached likelihood scoring on BANKING77 exited with status 250 after writing all 40 blocks of its eight cells; its blocks are retained and its cells classified as incomplete execution. The 30 one-factor cells are tabulated descriptively in Appendix J.

4.4 Batching versus shared-state reuse

The FP32 numerical prerequisite holds: all 12 same-precision path comparisons meet $\mathrm{atol} = 10^{-4}$ and $\mathrm{rtol} = 10^{-5}$ over 4,662 dependent request comparisons. BF16 does not: under looser tolerances

Metric (BANKING77 constrained, B=32)	Before	After
Batch-call latency, p50 (s)	1.233 [1.228, 1.252]	0.0936 [0.0926, 0.0948]
Throughput (decisions/s)	25.8 [25.6, 25.9]	342.9 [333.1, 344.5]
Gross GPU energy (J/decision)	4.30 [4.23, 4.38]	1.24 [1.22, 1.32]
Idle-adjusted GPU energy (J/decision)	0.961 [0.514, 0.993]	0.957 [0.952, 1.062]
Startup certification, 32 requests (s)	–	1.196

Table 2: Matched before/after trial on the fixed 32-request BANKING77 cohort (median and observed range over five blocks per visit). Before runs token-boundary certification inside every timed call. After runs it once at startup, outside the measured windows.

($\mathrm{atol} = 0.1$, $\mathrm{rtol} = 0.002$), only 1 of 12 same-precision and 0 of 16 BF16-versus-FP32 comparisons pass. Passing is neither bitwise equality nor general equivalence, and timed cells are FP32 only.

With $Q = 16$, fresh serial execution completes 12.3 and 11.8 decisions/s on the two states (Figure 2). Fresh batching raises this to 17.7 and 17.2, serial prefix reuse to 41.1 and 40.0, and parallel reused branches to 68.1 and 66.7, with small within-cell ranges. With $Q = 1$ all paths lie between 10.0 and 11.9 decisions/s, and the reuse paths are slower (10.0–10.4 versus 11.5–11.9), consistent with the prefix prefill and KV copy that reuse performs even for one question. Median sampled device peaks are 4.8 GB fresh serial, 8.2 GB fresh batched, 5.0 GB reused serial and 19.9 GB reused parallel (decimal units; the figure plots GiB). These cross-visit contrasts are descriptive and cover two states of about 8 kB.

4.5 Profile-guided host optimization

Moving certification to startup reduces median batch-call latency on the pre-certified cohort from 1.233 s to 0.0936 s and raises throughput from 25.8 to 342.9 decisions/s (Table 2); the visits ran about 91 s apart and passed the host filters. Certifying the 32 requests took 1.196 s. Adding this setup to the after median gives 1.290 s against 1.233 s before. This sum is cost accounting across separately measured windows, not a measured cold-call distribution, and it shows no demonstrated first-use reduction. Gross GPU energy falls from 4.30 to 1.24 J per decision, but idle-adjusted energy moves only from 0.961 to 0.957 J, below its 0.478 J floor, so no idle-adjusted energy reduction is established, and the lower gross energy does not by itself show less GPU work. First timed selections and statuses match in all five block pairs, yet logits differ in 3 to 31 of 32 requests per block (maximum absolute differences 0.33–0.74) and probabilities by at most 0.065. One request changes selection across repeated calls on both sides. The trial establishes no numerical equivalence and identifies no cause for the score differences.

5 Discussion and limitations

When eliminating generation pays off. On the tested workloads, bounded decisions did not need generated answers: the one-token path of the adapted decoder matched free generation from the same model to within 0.004 accuracy at the reported seed, and the fixed classifier, cross-encoder and likelihood scorers exceeded both on BANKING77 (Table 4). Eliminating generation did not, however, separate fast from slow paths. A one-token decoder already decides after one prefill, so an encoder head competes with that prefill while still encoding state, question and candidates, and host work was large enough to reorder costs. Prefix reuse and certification amortization paid off only under repetition. The one-factor cells suggest where the balance could shift (Table 8). Within each visit, batch size 32 raised query-head throughput 9.4-fold and decoder throughput 1.9-fold over batch size 1; a 2,048-token state cut the decoder from 164.0 to 66.2 decisions/s but the query head only from 46.8 to 38.1; and 2 or 32 candidates instead of 8 moved query-head throughput by less than its block range, while cached likelihood scoring fell from 94.9 decisions/s at 2 candidates to 30.6 at 32. At the constructed batch-32 anchor, the heads' medians (439 and 447) exceed the decoder's (317), the reverse of the natural cohorts. These cells use constructed requests, place decoders and heads in different sessions and register no cross-method direction or interaction, so they suggest, but do not establish, that the natural-cohort ordering depends on state length and batch size and need not hold elsewhere.

Implications for System One models. A model that returns typed decisions directly is one way to remove generation; a well-served one-token decoder is another, and it is the baseline an efficiency claim for parallel decision inference has to beat, on the same requests and with candidate encoding, host preprocessing, reuse and setup counted. Quality constrains which fast path is usable and depends on the population served, as the reversed classifier–decoder ranking on CLINC150 OOS queries shows, and candidate construction behaves as part of the predictor, as the constant abstainers show. We identified no architecture or kernel that maximizes quality per unit of latency and compute; the largest measured lever was host-side, and quality and cost remain separate estimands without a composite score. Because the interface and heads use only open weights, the comparisons can be repeated without proprietary System One models.

Limitations. The resource evidence covers two fixed, pre-certified cohorts on one GPU under monitored but uncontrolled host load. Decoders run in vLLM and encoders in PyTorch, in sessions about 24,000 s apart linked by a one-cell bridge at a different throughput scale; only decoder paths received the preprocessing optimization, decoder rates exclude certification, selections changed in 3 of 42 repaired cells and one visit is incomplete execution. No comparison therefore isolates architecture, separates budget from pretraining, establishes a quality/resource frontier or certifies quality preservation at batch size 32. Model scale varies only in released-decoder quality, precision only in the FP32/BF16 numerical checks, and no second GPU generation or multi-GPU topology is measured; GPU counters exclude whole-server energy, and human-labeled public English benchmarks do not establish behavior on live traffic. Changing-candidate results use one seed, one reordering per group and two regimes, and gold-included subsets measure selection among supplied alternatives, not retrieval. No hosted commercial System One model is benchmarked.

Future work. Interaction grids over context, candidate count, batch size, model scale, reduced precision, a second GPU generation and multi-GPU topology remain open, with cold-request measurements that include certification and matched encoder preprocessing. The measured fast paths and their cost accounting also frame two questions left open here: whether larger reasoning models can be distilled into them without losing calibration on changing candidate sets, and when a fast decision should be escalated to slower generative reasoning. Both need outcome-labeled workloads and quality-matched, multiplicity-controlled comparisons.

6 Conclusion

We asked whether bounded machine decisions need token-by-token generation, and when parallel structured decision inference improves latency and compute while preserving decision quality. On the tested workloads the decisions did not need generated answers, but whether a method generates answer tokens did not settle its cost. A one-token constrained decoder already avoids answer generation and outpaced our encoder query head on fixed, pre-certified batch-32 cohorts, and the largest measured improvement came from moving host-side certification out of repeated decoder calls, for requests prepared in advance. Quality and calibration depend on predictor, adaptation and workload, and candidate sets behave as part of the predictor. These are descriptive comparisons of tested implementations on one GPU; they neither isolate architecture nor certify quality at batch size 32. For System One models that serve as a fast decision layer inside software, the unit of comparison should be the complete decision, measured on shared requests and candidate sets with preprocessing, candidate encoding and setup counted and a well-served one-token decoder as a standing baseline. OpenSystemOne makes such comparisons repeatable with open weights, including for later work that distills slower reasoning models into fast decision paths or escalates uncertain decisions to them.

Ethics statement

This study evaluates decision inference on four public English benchmarks. We use them under the licenses declared by their upstream sources: BANKING77 (CC BY 4.0), CLINC150 (CC BY 3.0), SciQ (CC BY-NC 3.0, noncommercial) and BoolQ (CC BY-SA 3.0, share-alike). Their attribution and terms remain as declared, and we do not relicense the data. The anonymous reproduction package contains acquisition references, pinned revisions and hashes, model and data manifests, code, configurations, and results. It includes no raw benchmark text or model weights, so users obtain the

data from the original sources under the applicable terms. Lower inference cost and high benchmark accuracy do not by themselves make a method suitable for consequential automated decisions. Normalized candidate scores are not necessarily calibrated, and calibration must be checked on held-out data. Each application requires its own evaluation and appropriate human oversight.

REPRODUCIBILITY STATEMENT

Appendix A and Appendices E–K specify data splits, candidate conditions, training and selection, calibration, measurement protocol, placement, stopping rules and seeds, and Table 9 pins every public model revision. An anonymous artifact was prepared locally for this submission. It contains the code, pinned data-acquisition recipes, figure-reproduction scripts and unrounded numerical tables with source indices for every canonical endpoint and stratum, the changing-candidate contrasts and all 88 resource cells. Using the existing dependency environment, all 49 of its CPU interface checks were satisfied, and we regenerated both figures in a fresh directory with exactly matching PNG output. That check did not install dependencies anew or retrain models. Code and instructions are provided in the supplementary material.

AI USAGE DISCLOSURE

Generative AI tools were used to assist with retrieving and discovering related literature, code implementation, writing and polishing. The research ideas originated with the authors. All experimental results come from executed code and recorded outputs. The authors reviewed and verified AI-assisted content and take full responsibility for the paper's accuracy, originality, and conclusions.

REFERENCES

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient Intent Detection with Dual Sentence Encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, 2020. doi: 10.18653/v1/2020.nlp4convai-1.5. URL https://aclanthology.org/2020.nlp4convai-1.5/.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300/.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention. *arXiv preprint arXiv:2309.06180*, 2023. URL https://arxiv.org/abs/2309.06180.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. doi: 10.18653/v1/D19-1131. URL https://aclanthology.org/D19-1131/.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. In *International Conference on Machine Learning*, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Alexandra Sasha Luccioni, Yacine Jernite, and Emma Strubell. Power Hungry Processing: Watts Driving the Cost of AI Deployment? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024. doi: 10.1145/3630106.3658542. URL https://facctconference.org/static/papers24/facct24-6.pdf.

Chenxu Niu, Wei Zhang, Jie Li, Yongjian Zhao, Tongyang Wang, Xi Wang, and Yong Chen. TokenPowerBench: Benchmarking the Power Consumption of LLM Inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 32582–32590, 2026. doi: 10.1609/aaai.v40i38.40535. URL https://ojs.aaai.org/index.php/AAAI/article/view/40535.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. Leveraging large language models for multiple choice question answering. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.12353.

Ihor Stepanov, Mykhailo Shtopko, Dmytro Vodianytskyi, Oleksandr Lukashov, Alexander Yavorskyi, and Mykyta Yaroshenko. GLiClass: Generalist Lightweight Model for Sequence Classification Tasks, 2025. URL https://arxiv.org/abs/2508.07662. arXiv preprint.

Ihor Stepanov, Aleksandr Smechov, Mykhailo Shtopko, Dmytro Vodianytskyi, and Oleksandr Lukashov. SCX Router: Streaming Zero-Shot Model Selection with a Decoder-KV Classifier and a Real-World Task Ontology. *arXiv preprint arXiv:2609.02292*, 2026. URL https://arxiv.org/abs/2609.02292.

TheoLeeCJ. SemIf (formerly OpenJev), 2026. URL https://github.com/TheoLeeCJ/SemIf/tree/ca3ba65f142967030ecb453346e94d6f476a69df. Accessed September 24, 2026; pinned commit ca3ba65f142967030ecb453346e94d6f476a69df.

TypeSafe AI. System One, n.d. URL https://docs.typesafe.ai/concepts/system-one. Accessed September 19, 2026.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference, 2024. URL https://arxiv.org/abs/2412.13663. arXiv preprint.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin (eds.), *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4413. URL https://aclanthology.org/W17-4413/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Zihao Ye, Lequn Chen, Ruihang Lai, Wuwei Lin, Yineng Zhang, Stephanie Wang, Tianqi Chen, Baris Kasikci, Vinod Grover, Arvind Krishnamurthy, and Luis Ceze. FlashInfer: Efficient and Customizable Attention Engine for LLM Inference Serving. In *Proceedings of the 8th MLSys Conference*, 2025. URL https://arxiv.org/abs/2501.01005.

Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3914–3923, 2019. doi: 10.18653/v1/D19-1404. URL https://aclanthology.org/D19-1404/.

Urchade Zaratiana, Gil Pasternak, Oliver Boyd, George Hurn-Maloney, and Ash Lewis. GLiNER2: An Efficient Multi-Task Information Extraction System with Schema-Driven Interface. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 130–140, 2025. doi: 10.18653/v1/2025.emnlp-demos.10. URL https://aclanthology.org/2025.emnlp-demos.10/.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. SGLang: Efficient Execution of Structured Language Model Programs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 62557–62583, 2024. doi: 10.52202/079017-2000. URL https://arxiv.org/abs/2312.07104.

A Detailed decision and evaluation protocol

A.1 Decision interface and methods

A.1.1 Decision contract

A request is $r = (x, q, \mathcal{C})$ with state text x, question q and an ordered collection $\mathcal{C} = ((u_1, c_1), \ldots, (u_K, c_K))$ of stable candidate identifiers u_i and descriptions c_i. Targets are stored apart from requests and belong to the declared candidate space. Tasks that permit rejection include an explicit abstention candidate. The decision-only contract returns a selected identifier and a validity status. The full-distribution contract also returns one probability per candidate, computed from real-valued scores s_i as

$$p_i(T) = \frac{\exp(s_i/T)}{\sum_{j=1}^{K} \exp(s_j/T)}, \qquad T > 0. \tag{1}$$

Native probabilities use $T = 1$. Calibrated probabilities use a temperature fitted only on disjoint calibration data. An invalid or truncated generation is retained as a failed decision and is never replaced by a scorer output. A method that needs an extra scoring call to supply a distribution pays for that call in its full-distribution measurement.

A.1.2 Decoder paths

Free generation prompts with the state, the question and identifier–description pairs and requests only the chosen identifier. It decodes greedily with reasoning disabled, at most 32 new tokens and termination at the tokenizer's EOS token. Exact identifier parsing determines validity, and reaching the cap without EOS counts as truncation. Constrained generation restricts transitions to a trie of identifier strings followed by EOS. When every identifier is verified to be a single token at the actual prompt boundary, a separate one-token path reads the allowed-token logits from one prefill and returns the decision without a further decode step, so no EOS is decoded. Its candidate probabilities are allowed-token probabilities renormalized over the candidates. They are not probabilities of an unrestricted response. Multi-token identifier constraints form a separate condition.

Option likelihood scoring uses the tokenized prefix $P(x, q)$ of state and question. It treats each description followed by EOS as a continuation $y_i = (y_{i1}, \ldots, y_{i\ell_i})$:

$$s_i = \ell_i^{-\alpha} \sum_{t=1}^{\ell_i} \log p_\theta(y_{it} \mid P(x, q), y_{i,<t}), \qquad \alpha \in \{0, 1\}. \tag{2}$$

The raw sum ($\alpha = 0$) or token mean ($\alpha = 1$) is chosen on tuning data, and Equation 1 normalizes over candidates. Uncached scoring evaluates complete prefix–candidate sequences in candidate batches. Cached scoring computes the prefix once and evaluates suffixes from its KV state with the same weights, input tokens, termination and normalization. Every runtime measurement records the cache representation actually used and includes its materialization cost. Candidate descriptions sit in the generation prompt but form continuations in likelihood scoring. Generation versus likelihood scoring is therefore a method comparison, not a cache-only ablation (Appendix D).

Decoder requests must be tokenized so that the answer boundary preserves the intended prefix tokens. We call this check token-boundary certification. In the measured decoder paths, certification

Population	Train	Tune	Calibrate	Test	Test groups	K
BANKING77	8,053	973	966	3,079	3,079	77
CLINC150, in scope	14,995	1,516	1,482	4,498	4,498	151
CLINC150, OOS	100	44	56	1,000	1,000	151
SciQ, with support	10,427	470	414	879	879	4
SciQ, missing support	1,191	57	56	111	111	4
BoolQ	6,404	804	780	3,270	2,938	2

Table 3: Retained examples after preprocessing for the known-inventory task splits. Test groups are the original text or passage units used for uncertainty estimates; permutations do not increase this count. K is the canonical candidate count. CLINC150 includes an explicit abstention candidate. SciQ support strata are kept separate, and BoolQ can contain multiple questions per passage group.

runs once per request at startup, outside timed calls, and its elapsed time is reported separately. Timed calls then tokenize the prefix once per request. Likelihood scoring still tokenizes each candidate continuation anew, which amounts to $K+1$ tokenizer calls per request, and caches no candidate tokens. The constrained and generation paths cache only identifier and certification identities, never prompts or predictions. In certified mode a changed or unseen request is rejected until it has been prepared, and the default serial mode remains available. Resource measurements of decoder paths use vLLM (Kwon et al., 2023) in BF16. Encoder paths run in our PyTorch implementation with BF16 autocast, and their tokenization stays inside timed calls.

A.1.3 Encoder paths and learned heads

The fixed classifier pools an encoding of state and question, applies a linear head over the known label inventory and normalizes over the requested subset. It has no output for a new label. The cross-encoder scores each state–question–candidate triple with a shared scalar head. This conditions the state representation on the candidate at the price of repeating state processing for every candidate. The dual encoder pools state–question and candidate encodings separately and scores their normalized projections with a learned-scale dot product. Candidate encoding counts toward the cost of every measured request.

The query head encodes the state once into token representations $H \in \mathbb{R}^{L\times d}$ and pools separate question and candidate encodings into h_q and h_{c_i}. With shared projections $a = W_q h_q$ and $b_i = W_c h_{c_i}$,

$$z_i = \mathrm{LN}\big(a + b_i + \mathrm{MHA}(a + b_i, W_x H, W_x H)\big), \qquad s_i = \mathrm{MLP}\big([z_i; b_i; a]\big). \tag{3}$$

All candidate queries attend to the same state memory in one batched call. The set head adds residual self-attention over $z_1, \ldots, z_K$ before scoring and uses no candidate-index embeddings. Reordering candidates therefore reorders scores in exact arithmetic; tied decisions and finite-precision effects are examined separately. The pooled-query control replaces the state-token memory with its masked mean. It separates the value and cost of token-level state access from candidate interaction. The heads instantiate a public typed-decision interface, which is the sense in which we call them Jev-like (TypeSafe AI, n.d.). They do not reproduce proprietary Jev weights, architecture or training.

A.2 Experimental design

A.2.1 Workloads and candidate sets

The suite comprises BANKING77 intent routing, CLINC150 intent and out-of-scope classification (Casanueva et al., 2020; Larson et al., 2019), SciQ answer selection (Welbl et al., 2017) and BoolQ passage-conditioned yes/no verification (Clark et al., 2019). Test evaluation uses the official labeled test splits, and BoolQ uses its labeled validation split. Normalized duplicate text and shared passages define the leakage groups that keep training, tuning, calibration and test disjoint, and each task model is trained separately. Table 3 gives the retained populations. Its test groups are the resampling units.

Canonical evaluation uses every retained test example with its original candidates and is the primary quality population. CLINC150 out-of-scope queries form a separate rejection stratum. SciQ examples with originally missing support are separated from supported ones. A stable 60/20/20 hash

partition of semantic labels yields held-out-label test views of 520 BANKING77 examples over 13 labels and 1,018 CLINC150 examples over 34 labels. The fixed classifier does not apply to these views. SciQ keeps its natural four-option task and BoolQ remains binary (Appendix E).

A.2.2 Comparison regimes and coverage

Released checkpoints and task-adapted models form separate regimes. Released endpoints have no training seed; they include Qwen3-0.6B (Yang et al., 2025) and the GLiClass-modern checkpoint (Stepanov et al., 2025). Adapted models build on Qwen3-0.6B and ModernBERT-base (Warner et al., 2024). They share examples, candidate construction and ordering seeds, and they follow a fixed training budget with tuning-only selection (Appendix F). The budgets make supervision explicit without equalizing adaptation capacity across architectures.

Twenty-six primary families are trained at seeds 17, 29 and 43. They comprise fully adapted ModernBERT-base query and set heads for every task and label regime, the applicable fully adapted fixed classifiers, and Qwen3-0.6B adapted with LoRA under a candidate-normalized likelihood objective and under a generation objective. Of the 70 supporting single-seed families, 16 are evaluated. Twelve cover BANKING77 known labels: nine ModernBERT-base configurations with full or frozen backbones and three fully adapted ModernBERT-large configurations. The other four are the fully adapted cross-encoder and pooled-query head on SciQ and BoolQ. Released Qwen3-1.7B and Qwen3-4B and a second GLiClass backbone are evaluated on BANKING77 known labels as bounded quality checks of scale and backbone. The Qwen3 checkpoints cover constrained, likelihood and cached likelihood scoring, and none of these checks enters a resource measurement. Together these give 167 primary canonical endpoints. Each endpoint is a fixed method, dataset and label regime, checkpoint and runtime evaluated at batch size 1. Twenty-two further sensitivity endpoints rescore likelihood paths whose tuning-selected rule is the token mean with the raw log-probability sum instead. Coverage of the seven families is therefore specific to where each applies and does not form a complete method-by-dataset product. Single-seed configurations are never reported as replicated.

A.2.3 Canonical quality and calibration

Accuracy and macro F1 summarize decisions. Invalid and truncated outputs stay in the attempted-example denominator and are also reported as coverage. For full-distribution outputs, NLL and the multiclass Brier score are the primary probability metrics. Expected calibration error with 15 equal-width bins is a diagnostic. No probability metric is imputed for decision-only output. A positive scalar temperature (Guo et al., 2017) is fitted by NLL on the disjoint canonical calibration split only, and native and calibrated scores are both reported (Appendix G). Descriptive 95% intervals use 2,000 cluster-bootstrap resamples over original test groups, which preserve shared BoolQ passages. Trained seeds are reported separately. For intent likelihood training, the seed-29 and seed-43 replicas differ from seed 17 in one recorded execution setting (Appendix F). Seed spread is therefore observed replicate variation, not a pure estimate of seed variance. No multiplicity-adjusted superiority claim is made.

A.2.4 Changing candidate sets

Two task and label regimes probe changing candidates at checkpoint seed 17. The first is BANKING77 with held-out labels, whose test inventory has 13 labels. The second is CLINC150 with known labels, which includes out-of-scope queries. Seven method identities are evaluated in each regime: adapted free generation, adapted constrained decoding, adapted uncached and cached likelihood scoring, the fully adapted query and set heads, and released GLiClass-modern. This gives 14 inference jobs at batch size 32. Groups are selected by a salted hash of the group identifier within declared population quotas, without using model outcomes. The quotas are 128 BANKING77 held-out-label groups and, for CLINC150, 96 in-scope and 32 out-of-scope groups (Appendix E). Every method receives identical requests.

BANKING77 has eight conditions on all 128 groups. Two are the canonical candidate set in its original order and under one seeded permutation. The others are gold-included and gold-absent subsets with $K \in \{2, 8, 13\}$. CLINC150 has the two canonical conditions on all 128 groups, gold-included and gold-absent subsets with $K \in \{2, 8, 32\}$ on the 96 in-scope groups, and out-of-scope conditions

with $K \in \{2, 8, 32\}$ on the 32 out-of-scope groups. The subset conditions contain the abstention candidate `__abstain__`, and the canonical BANKING77 set does not. Equal K therefore need not mean equal candidates. Gold-absent and out-of-scope requests have abstention as their target, so their accuracy measures abstention. Each endpoint receives 1,024 BANKING77 or 928 CLINC150 requests, 13,664 in total.

Each condition is compared with the unpermuted canonical condition at the same batch size. The comparison is restricted to exactly that condition's groups and uses paired group-clustered bootstrap intervals. CLINC150 in-scope and out-of-scope contrasts never use the pooled 128-group reference. Requests run condition by condition in one frozen group order and contiguous batches of 32. The permuted and unpermuted canonical conditions therefore share batch companions and differ only in candidate order. The other conditions also change the candidate sets of companion requests and, in CLINC150, which groups share a batch. Their contrasts combine candidate and batch-composition changes and are not interpreted as purely caused by the candidates. Permutation results describe one fixed reordering per group. They make no claim of permutation invariance, variability across orderings or generality across training seeds. Calibration transfer applies each endpoint's temperature, fitted once on its canonical calibration split at batch size 1, unchanged to every condition. No temperature is fitted on test variants. Uncached and cached likelihood scoring are treated as two approximate BF16 predictors, with neither equivalence nor a cache benefit claimed.

A.2.5 Controlled resource measurements on one Blackwell GPU

The server has four NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs connected over PCIe in two NUMA affinity groups, with no NVLink observed. All resource and optimization measurements run on one of these GPUs. Its physical identity is verified against a device query before every runtime visit, and every method and cell uses the same CPU affinity. Jobs of other users on other GPUs are neither paused nor terminated. We record and filter host contention and do not claim isolation (Appendix H).

The 88 cells are of three kinds. Natural-input cells measure 21 endpoint identities at batch sizes 1 and 32 (42 cells) on two 32-request test cohorts fixed by request identifier before collection. These are BANKING77 known-label in-scope requests and SciQ passage-conditioned requests. The BANKING77 identities are the following: adapted free generation, constrained decoding, and uncached and cached likelihood scoring; the fully adapted fixed classifier, cross-encoder, query head and set head; released GLiClass-modern; and released uncached and cached Qwen3-0.6B likelihood scoring. SciQ has the same set without the fixed classifier. One-factor cells (30) run five methods: adapted constrained decoding, released uncached and cached scoring, and the query and set heads. They use constructed BANKING77-derived long-context requests at six coordinates of (state tokens, K, batch size). These are the anchor $(512, 8, 1)$, a context change to $(2048, 8, 1)$, cardinality changes to $(512, 2, 1)$ and $(512, 32, 1)$, and batch changes to $(512, 8, 8)$ and $(512, 8, 32)$. Each contrast changes one factor and estimates no interaction. The constructed requests keep an artificial context boundary, so these cells carry cost evidence without a semantic accuracy claim. The 16 shared-state cells are described in Section A.2.6.

Every natural and one-factor cell cycles through the same 32 requests in file order. At batch size 1 a cycle is 32 serial single-request calls, at batch size 8 it is four calls of eight, and at batch size 32 it is one call. Each cell has five measurement blocks. Each block has an excluded warmup call, minimum and maximum active times and a matched idle measurement (Appendix H). Host timing uses a monotonic clock with device synchronization at call boundaries. It covers tokenization, transfer, inference, normalization and output assembly. Model loading, compilation, warmup and decoder startup certification fall outside the measured windows. Throughput is completed decisions per active second. At batch size 1 each call carries one request, so call latency equals request latency. At batch sizes 8 and 32 we report batch-call latency, which is never called request latency. Energy is the difference of cumulative GPU energy counters over the active window, reported as gross and idle-adjusted joules per decision. CPU, host-memory and power-supply energy lie outside this boundary. The collector samples device memory occupancy and records parent-process allocator peaks. Selected identifiers and statuses are saved per request and call. Block summaries describe execution variability on declared inputs, not population uncertainty.

Collection ran in two sessions. The original session collected the 25 runtime visits in seeded random order: 21 endpoint identities and four shared-state paths. Each identity was loaded once with all its

cells, and each cell had five blocks in recorded randomized order. After the optimization trial of Section A.2.7, the 12 decoder visits were collected again with certification at startup. The final census combines the 13 original encoder and shared-state visits with the 12 repaired decoder visits. It comprises 25 visits, 88 cells, 100 registered contrasts and 440 blocks, all of them clean under the contamination rule below. One repaired visit, released Qwen3-0.6B uncached likelihood scoring on BANKING77, exited with status 250 after writing all 40 blocks of its eight cells. Its blocks are retained, and its cells are classified as incomplete execution. The superseded original decoder collections remain in the artifact and serve as the reference for decision comparisons.

A block is contaminated when a foreign process appears on the measured GPU or a required telemetry sample is missing. Contaminated blocks are excluded from clean summaries and are not replaced. Contrasts within one visit compare blocks from the same model load. Contrasts across visits include every comparison between methods and between shared-state paths. They are interpreted only when both visits pass host-load filters. A direction is stated only when the difference of block medians exceeds the larger within-visit block range. Contrasts between an original-session visit and a repaired-session visit face a further screen based on one bridge cell. The bridge cell is the BANKING77 batch-32 constrained cell with certification inside timed calls, measured in both sessions. A direction is retained only when the contrast's difference exceeds the bridge's between-session shift in the same metric. It is suppressed when the bridge lacks that metric. Batch-32 bridge calls have no request latency, so every cross-session batch-1 request-latency direction is suppressed. Batch-1 call latency equals request latency in value, so we treat both as descriptive. These filters reduce between-visit shifts from reloads, allocator state, clocks or temperature, but they neither identify nor bound them. All resource comparisons are therefore descriptive and conditional.

A.2.6 Batching versus state reuse

For RQ3, a bounded shared-state experiment poses repeated questions about the same authored state. Four HF/PyTorch paths form a 2×2 design: fresh serial, fresh batched, reused-prefix serial and reused-prefix parallel. All four use released Qwen3-0.6B in FP32, the same prompts and the same single-token candidate-probability contract. Bundles of the first $Q \in \{1, 16\}$ questions in original order run on two fixture states. These are the shortest and the median of the 37 states by UTF-8 byte length, giving 16 cells. Complete calls include tokenization, KV copying and materialization, padding and output assembly. Each call returns Q decisions, and blocks follow the same stopping rules. The four paths are separate runtime visits, so reuse contrasts are cross-visit. Measured outputs are checked against FP32 numerical tolerances before any interpretation as matched computation. The fixture is unlabeled and its states span only 8,004–8,229 bytes. Conclusions therefore concern implementation costs on these two states. They give no semantic-quality evidence, say nothing about longer states or length scaling and do not estimate a vLLM cache gain (Appendix I).

A.2.7 Profiling and one matched optimization

Profiling targets host inference overhead in two natural BANKING77 cells at batch size 32: adapted constrained decoding and the fully adapted query head. It uses the measurement GPU and CPU affinity. After one warmup, Python cProfile records 10 identical calls per endpoint with device synchronization at call boundaries; these runs are not resource measurements. Eligible components are those to which the profile attributes host time and whose removal leaves the defined computation unchanged. Changes to the model, weights, precision, prompt, candidate count, batch size or decision rule are excluded. Components are ranked by exclusive self time summed over the 10 calls. cProfile measures this self time in wall-clock seconds, not summed CPU-core time. It does not observe GPU kernels, so the optimization supports no kernel-level claim. The selected component was the per-request token-boundary certification that the decoder paths ran inside every timed call.

The matched trial runs two separate visits in fixed order, before and then after. Both use the same GPU, cell, requests, checkpoint, precision and affinity, and they differ only in whether certification runs at startup. Each visit has five blocks. For correctness we pair block k of the two visits by the first timed observation of each request. We compare selected identifiers and statuses, report score and probability differences without tolerance and record whether repeated calls of a request change its selection. This reports disagreement and is not an equivalence certificate. The after configuration was then applied to all decoder paths. The query head and the other encoder paths were left unchanged.

A.2.8 WHAT EACH COMPARISON ATTRIBUTES

Encoder-versus-decoder differences combine pretrained capability, tokenization, serving engine and execution. Cached versus uncached scoring holds the model and the scoring rule fixed, but in BF16 the two paths are distinct approximate predictors. The shared-state paths hold the checkpoint, prompts and probability contract fixed and vary only batching and reuse. Query versus set heads, and token-level versus pooled state, compare learned variants trained on the same data. The before/after optimization changes one host-side component of an otherwise fixed path.

B EXPANDED QUALITY, DYNAMIC-OPTION AND SERVING RESULTS

	BANKING77	CLINC150		SciQ		BoolQ
Endpoint		in scope	OOS	passage	missing support	
n	3,079	4,498	1,000	879	111	3,270
Adapted generation	0.805	0.886	0.905	0.994	0.793	0.812
Adapted constrained	0.804	0.886	0.909	0.994	0.793	0.812
Adapted likelihood	0.864	0.942	0.784	0.990	0.748	0.801
Adapted cached likelihood	0.864	0.942	0.782	0.990	0.748	0.800
Fixed classifier	0.905	0.955	0.730	–	–	–
Cross-encoder	0.888	–	–	0.994	0.559	0.761
Query head	0.733	0.862	0.718	0.474	0.550	0.658
Set head	0.648	0.749	0.844	0.667	0.468	0.642
Released GLiClass-modern	0.379	0.450	0.000	0.923	0.550	0.553
Released Qwen3-0.6B constrained	0.203	0.213	0.000	0.917	0.631	0.633
Released Qwen3-0.6B likelihood	0.131	0.002	1.000	0.810	0.369	0.383

Table 4: Canonical accuracy on the retained evaluation examples (BoolQ: official validation split) in the known-label regimes, at seed 17 for adapted endpoints. Adapted decoders are Qwen3-0.6B with LoRA, and adapted encoder paths are fully adapted ModernBERT-base. Invalid and truncated outputs count as errors, and a dash marks a configuration not evaluated or not applicable. The released likelihood scorer's CLINC150 accuracies (0.002 in scope, 1.000 OOS) reflect selecting abstention for nearly every request. The artifact's canonical table holds the remaining values: grouped 95% intervals, macro F1, NLL, ECE, temperatures, seeds 29 and 43, and held-out-label regimes.

BANKING77, held-out labels, 128 groups						
Method	Canonical	Perm. changes	Included K=2	Included K=13	Absent K=2	Absent K=13
Adapted generation	0.836	9	0.953	0.805	0.789	0.398
Adapted constrained	0.836	9	0.945	0.805	0.805	0.414
Adapted likelihood	0.367	0	0.000	0.000	1.000	1.000
Adapted cached likelihood	0.367	2	0.000	0.000	1.000	1.000
Query head	0.266	0	0.727	0.266	0.750	0.125
Set head	0.117	0	0.969	0.000	0.086	1.000
Released GLiClass-modern	0.625	26	0.805	0.594	0.688	0.250
CLINC150, known labels, 96 in-scope and 32 OOS groups						
Method	Canonical	Perm. changes	Included K=32	Absent K=32	OOS canonical	OOS K=32
Adapted generation	0.844	7 / 3	0.885	0.948	0.938	0.938
Adapted constrained	0.833	6 / 3	0.885	0.948	0.938	0.938
Adapted likelihood	0.906	1 / 0	0.938	0.938	0.750	0.969
Adapted cached likelihood	0.906	0 / 0	0.938	0.927	0.750	0.969
Query head	0.823	0 / 0	0.875	0.865	0.656	0.969
Set head	0.708	0 / 0	0.885	0.792	0.906	0.969
Released GLiClass-modern	0.427	43 / 18	0.604	0.208	0.000	0.406

Table 5: Changing-candidate accuracy at seed 17 and batch size 32. Canonical is the unpermuted canonical set on the selected groups (in scope for CLINC150). Perm. changes counts requests whose selected identifier changes under one seeded reordering (CLINC150: in scope / OOS). Included and Absent are gold-included and gold-absent subsets of size K. These subsets contain the abstention candidate, which the canonical BANKING77 set lacks. Gold-absent and OOS accuracy is the abstention rate. All conditions are dependent views of the same groups. The artifact holds paired group-clustered intervals, NLL before and after the reused canonical temperature, and the K=8 conditions.

	BANKING77 cohort		SciQ cohort	
Endpoint identity	Decisions/s, B=32	Test acc.	Decisions/s, B=32	Test acc.
Adapted generation	342.5 [333.6, 348.1]	0.805	880.0 [851.5, 888.2]	0.994
Adapted constrained	342.3 [276.1, 347.3]	0.804	897.3 [888.6, 906.2]	0.994
Adapted likelihood	31.1 [30.6, 33.3]	0.864	199.2 [187.0, 207.1]	0.990
Adapted cached likelihood	36.1 [33.9, 36.9]	0.864	399.9 [391.5, 419.9]	0.990
Fixed classifier	2842.2 [2817.2, 2919.8]	0.905	–	–
Cross-encoder	49.3 [49.1, 49.6]	0.888	144.0 [143.4, 144.2]	0.994
Query head	224.4 [219.0, 244.9]	0.733	582.3 [578.6, 588.4]	0.474
Set head	238.3 [214.4, 246.4]	0.648	575.8 [570.0, 577.2]	0.667
Released GLiClass-modern	819.6 [812.1, 828.6]	0.379	739.8 [733.7, 746.0]	0.923
Released likelihood†	32.2 [31.5, 33.2]	0.131	190.4 [186.8, 199.1]	0.810
Released cached likelihood	36.0 [35.7, 38.4]	0.130	389.8 [326.4, 398.9]	0.805

Table 6: Batch-32 throughput on the fixed 32-request cohorts (median and observed range over five blocks) beside canonical full-test accuracy at batch size 1 on the corresponding population (BANKING77 known labels; SciQ passage-conditioned). The two columns are separate estimands. Released decoders are Qwen3-0.6B. †BANKING77 cell from the visit classified as incomplete execution. Directions are stated only for the registered contrasts discussed in the text.

C Paired comparisons with trained baselines

Section 4.1 summarizes six paired comparisons between the adapted constrained decoder and fully fine-tuned ModernBERT-base encoders. The decoder is Qwen3-0.6B with LoRA, evaluated through the one-token constrained path in vLLM in BF16. The encoders run in our PyTorch implementation with BF16 autocast. Each comparison fills one role. Constrained decoding is the optimized generation reference from RQ1, the fixed classifier is a cheap baseline for a known label inventory, and the cross-encoder conditions explicitly on each candidate. The fixed classifier is compared with the decoder on BANKING77 and on the CLINC150 in-scope and OOS strata. The cross-encoder is compared with the decoder on BANKING77, BoolQ and SciQ passage-conditioned questions. The selection is not exhaustive. Likelihood scorers, query and set heads, and released or larger backbones remain in Tables 1 and 4, and SciQ missing-support questions are not paired here.

We chose these comparisons after the per-side accuracies were visible, including those of the seed-29 and seed-43 replicas. The comparison set was declared before the paired bootstrap was run, but it is neither blinded nor a preregistered replication. Every available replica of the selected roles is included regardless of direction, which guards against omitting unfavorable seeds. The three classifier roles have trained replicas at seeds 17, 29 and 43. The three cross-encoder roles exist only at seed 17, so there are 12 seed-specific paired rows. For each family, the tuning-data search at seed 17 set the learning rate. The replicas reuse that configuration, and their recorded training configurations differ from seed 17 only in the seed and output directory. Each replica selects its own checkpoint epoch on tuning data. Selected epochs agree across seeds except for the seed-43 CLINC150 constrained decoder, which selected zero-based epoch 1 rather than 2. Equality of execution beyond the recorded configurations is not established.

Each row uses every request of its stratum in the retained evaluation population at batch size 1 (for BoolQ, the official validation split), and both systems score identical requests. Differences are encoder minus decoder. Invalid outputs would stay in the accuracy denominator, but no row contains one. NLL and Brier score use native probabilities ($T = 1$) without refitting, and both systems returned finite probabilities for every paired request. Each row has its own 2,000 bootstrap resamples, drawn with seed 1701 over entire original groups. This keeps the 3,270 BoolQ requests clustered in 2,938 passage groups. None of the 36 marginal 95% intervals (three metrics in each of 12 rows) is adjusted for multiplicity. A seed-17 row conditions on one trained model per side, so its interval reflects evaluation sampling and not training-seed variation. The three rows of each classifier role are neither pooled nor averaged. Their agreement in direction describes these observed pairs, not a population of training seeds.

Backbone, pretraining, adaptation method and serving runtime differ within every pair. The rows therefore compare these trained systems and attribute no difference to architecture. They neither

show that classifiers generally outperform decoders nor identify a best family. The fixed classifier has no output for labels outside its training inventory, and the cross-encoder results do not establish transfer to unseen labels. On SciQ both systems reach 0.994. The zero difference and its ± 0.68-point interval therefore reflect a ceiling and are not evidence of equivalence. These batch-1 quality contrasts are a separate estimand from the batch-32 cohort costs of Section 4.3. They define no quality-matched speedup or frontier.

(a) Accuracy

Comparison	Seed	n	Enc.	Dec.	Δ accuracy (pp)
Classifier, BANKING77	17	3,079	0.905	0.804	$+10.17$ [8.80, 11.46]
	29	3,079	0.922	0.808	$+11.33$ [10.00, 12.67]
	43	3,079	0.912	0.818	$+9.35$ [8.12, 10.65]
Classifier, CLINC150 in	17	4,498	0.955	0.886	$+6.91$ [6.00, 7.80]
	29	4,498	0.947	0.881	$+6.67$ [5.76, 7.58]
	43	4,498	0.955	0.857	$+9.83$ [8.80, 10.89]
Classifier, CLINC150 OOS	17	1,000	0.730	0.909	-17.90 [-20.80, -15.00]
	29	1,000	0.707	0.914	-20.70 [-23.60, -17.90]
	43	1,000	0.710	0.841	-13.10 [-16.20, -10.20]
Cross-encoder, BANKING77	17	3,079	0.888	0.804	$+8.41$ [7.08, 9.74]
Cross-encoder, BoolQ	17	3,270	0.761	0.812	-5.08 [-6.80, -3.45]
Cross-encoder, SciQ passage	17	879	0.994	0.994	0.00 [-0.68, 0.68]

(b) Native proper scores

Comparison	Seed	Δ NLL	Δ Brier
Classifier, BANKING77	17	-0.551 [-0.614, -0.492]	-0.147 [-0.164, -0.129]
	29	-0.556 [-0.615, -0.500]	-0.160 [-0.177, -0.143]
	43	-0.491 [-0.547, -0.442]	-0.136 [-0.153, -0.120]
Classifier, CLINC150 in	17	-0.499 [-0.559, -0.439]	-0.118 [-0.132, -0.104]
	29	-0.465 [-0.523, -0.408]	-0.115 [-0.128, -0.101]
	43	-0.557 [-0.617, -0.501]	-0.151 [-0.165, -0.137]
Classifier, CLINC150 OOS	17	$+0.346$ [0.229, 0.462]	$+0.286$ [0.247, 0.325]
	29	$+0.345$ [0.236, 0.460]	$+0.304$ [0.265, 0.342]
	43	$+0.250$ [0.139, 0.368]	$+0.220$ [0.180, 0.263]
Cross-encoder, BANKING77	17	-0.511 [-0.569, -0.455]	-0.125 [-0.141, -0.109]
Cross-encoder, BoolQ	17	$+0.096$ [0.066, 0.127]	$+0.062$ [0.042, 0.082]
Cross-encoder, SciQ passage	17	-0.001 [-0.016, 0.012]	-0.0002 [-0.008, 0.007]

Table 7: Six post-hoc role comparisons in 12 seed-specific paired rows on identical requests at batch size 1. Differences are encoder minus decoder. Encoders are fully fine-tuned ModernBERT-base, and the decoder is Qwen3-0.6B with LoRA and one-token constrained decoding. n counts paired requests, all with finite native probabilities on both sides. The BoolQ requests (official validation split) form 2,938 passage groups, and every other request is its own group. Enc. and Dec. give accuracy. Accuracy differences are in percentage points. Positive accuracy differences and negative NLL or Brier differences favor the encoder. Brackets are unadjusted marginal 95% intervals from 2,000 whole-group bootstrap resamples (seed 1701). Rows are not pooled across seeds.

D METHOD IMPLEMENTATION DETAILS

Request and candidate identifiers remain stable when outputs are mapped back to evaluation records. For likelihood scoring, EOS is included in both the sum and the length of Equation 2. Token-boundary certification tokenizes every candidate string together with the prefix. It verifies that the answer boundary preserves the intended prefix token sequence, and the same check confirms single-token identifiers for the one-token constrained path. In the measured decoder paths this certification runs at startup for every measured request. Its elapsed time and the number of newly certified re-

quests are recorded per visit and cell. A request certified earlier has no new setup cost, which the artifact records as a null per-request cost, never as a zero first-use cost. The HF reference cached implementation copies and repeats prefix KV for candidate batches and does not provide zero-copy sharing. Prefix reuse removes repeated prefix computation. Candidate suffixes, vocabulary projections, KV handling and normalization remain, so we measure these costs and do not convert a reduction in prefix evaluations into an asserted speedup.

The dual encoder L_2-normalizes its projections. Its candidate embeddings are reusable only when text, model and execution settings match. In Equation 3 the notation suppresses projection biases and the conventional orientation of token matrices. The query head uses projected width 256, four attention heads, a residual connection, layer normalization and a shared two-layer GELU scorer. In the set head every projection and scoring function is shared across candidates, which yields permutation equivariance in exact arithmetic.

E WORKLOADS AND CANDIDATE CONDITIONS

When overlap requires removing rows, evaluation takes priority over calibration, which takes priority over tuning and then training. Five utterances occur in both BANKING77 and CLINC150, and each task model is trained only on its own dataset.

The semantic label partition assigns training, tuning and test label groups by a stable 60/20/20 hash. Calibration uses calibration examples whose labels lie in the tuning partition, so test performance measures transfer to semantic labels unseen in training and calibration. The held-out-label views add no independent examples to the canonical test populations, and statements about them refer to this partition only. Transfer to other unseen labels is not estimated.

The changing-candidate groups are drawn from a previously fixed cohort. For each dataset and label regime, that cohort keeps the 512 original test groups with the lowest seed-1701 hash of their group identity, or every group when fewer exist. Within that cohort and within each population (BANKING77 held-out labels; CLINC150 known in-scope; CLINC150 known out-of-scope), groups are ranked by the SHA-256 digest of the string `scope-v10:1701:` followed by the group identifier. Groups are taken in ascending order of this digest until the quotas of 128, 96 and 32 groups are filled. The rankings use group identity only, never predictions, labels, correctness or calibration outcomes. The selected groups are then merged and sorted by the same digest and then by group identifier. Collection is condition-major: each condition keeps this relative order restricted to its eligible groups and is split into contiguous batches of 32. Every condition uses permutation seed 1701. Before collection, each unpermuted canonical request is checked against the canonical batch-1 request for the same group. A mismatch is treated as an input preparation error, never as a reason to drop a group.

Per endpoint, BANKING77 contributes $8 \times 128 = 1{,}024$ requests and CLINC150 contributes $128 + 128 + 6 \times 96 + 3 \times 32 = 928$. The 14 jobs therefore contain $7 \times 1{,}024 + 7 \times 928 = 13{,}664$ requests. Repeated conditions of one group are correlated observations of the same original example, and all uncertainty is clustered by group.

F TRAINING, SELECTION AND REPLICATION

Intent training varies candidate sets across epochs. For in-scope examples it removes the correct label with deterministic probability 0.25, and abstention appears in both the included and absent conditions; candidate sampling uses seed 1701. Every run uses three epochs, effective batch size 32, AdamW with weight decay 0.01 and gradient clipping at 1. Full encoder adaptation searches backbone learning rates $\{10^{-5}, 3 \times 10^{-5}\}$ and head rates $\{10^{-4}, 3 \times 10^{-4}\}$. Frozen-backbone conditions search only the head rates. LoRA uses rank and alpha 16, zero dropout, all-linear targets and the two backbone rates. Likelihood adaptation minimizes cross-entropy after normalizing complete candidate scores. Generation adaptation minimizes the full-vocabulary NLL of the gold identifier and EOS, which is not interchangeable with candidate-normalized NLL. Larger pinned models receive no training search.

The seed-17 search covers 276 training cells in 96 dataset/model/regime families. Configuration and epoch are selected per family on tuning data alone, with deterministic tie-breaking over epochs

and learning rates. A run is eligible when all three tuning epochs completed with finite values, its selected epoch is consistent, and it recorded no timeout, interruption, training error or artifact error. Incomplete auxiliary energy telemetry does not exclude an otherwise complete run. That run's training energy is reported as unavailable, not imputed, and runs are not repeated only to recover it.

The 26 primary families are retrained from scratch at seeds 29 and 43 with their selected configurations, adding 52 runs. A replica enters analysis only after a checkpoint census confirms its completion and provenance. An incomplete replica remains unavailable and is never silently replaced or rerun wholesale. For intent likelihood training, the seed-29 and seed-43 runs evaluate candidates in chunks of 64, and the seed-17 runs use chunks of four. In BF16 this changes the gradient arithmetic, so differences between seed 17 and these replicas combine training randomness with this recorded execution difference. Paired analyses compare matching trained seeds and do not pool seed replicas as independent task examples. All search and replication costs are retained, including failed and interrupted runs. Test labels select no prompt, normalization, checkpoint or optimization path.

G Normalization and calibration rules

The multiclass Brier score is summed over candidates and averaged over examples. Expected calibration error uses 15 equal-width bins, with 10- and 20-bin sensitivity. The choice between raw-sum and length-normalized likelihood scores is made on tuning data, and uncached and cached likelihood endpoints of one model share the selected rule. The 22 sensitivity endpoints show that the rule matters on test data. For example, rescoring the seed-17 adapted likelihood scorer on BANKING77 held-out labels with the raw sum gives accuracy 0.456, against 0.363 for the tuning-selected token mean. The temperature is a positive scalar fitted by calibration NLL on the canonical calibration split, independently of test outcomes. Four primary fits reach the fitting boundary $T = 20$. They are the seed-29 set head on CLINC150 held-out labels, the seed-43 set head on BANKING77 held-out labels, and released uncached and cached Qwen3-0.6B likelihood scoring on BoolQ. For held-out labels, the fit uses calibration examples with tuning-partition labels and never test label identities. The changing-candidate conditions reuse each endpoint's canonical batch-1 temperature unchanged. Calibration under changed candidate sets is therefore a transfer estimate, and no temperature is fitted on test variants or on new calibration collections.

H Systems measurement protocol

Device and host placement. The measured GPU is physical GPU 2, attached to NUMA node 1. Before each visit the collector queries the device's physical index, UUID and PCI bus address. It saves the raw query with the visit record and refuses to collect on a mismatch with the expected identity. Measured processes are pinned to logical CPUs 48–95. During each block, telemetry records compute processes on the measured GPU, activity on GPUs 0, 1 and 3, and host CPU load on NUMA node 1. A process that starts and exits between samples can escape detection, so exclusive device access is not claimed.

Blocks. A measurement block starts with one excluded warmup call. It then runs whole cycles until, at a cycle boundary, it has at least 10 measured calls, one complete cycle and 2 s of active time, or until active time reaches 30 s. The cap never cuts a cycle short, and overruns are recorded. A block that the cap stops before the minima are met is flagged as cap-limited and disclosed, not rerun. A 2 s idle measurement with the model resident accompanies each block. The original-session visit order is a shuffle, seeded with 1701, of the lexicographically sorted runtime identities.

Cell accounting. Consider a block with C completed 32-request cycles, active time T and GPU energy E over the cycle-aligned active window. Its throughput is $32C/T$ and its gross energy per decision is $E/(32C)$. The idle-adjusted value subtracts the duration-scaled energy of the block's matched idle measurement and can be negative. Shared-state cells count Q decisions per call and stop after at least 10 calls and 2 s of active time, with the same 30 s cap. A failed call, out-of-memory event, unsupported configuration or incomplete cycle is retained as an execution failure and never counted as a completed cycle. A visit whose process exits abnormally keeps its saved blocks and is classified as incomplete execution, and no exit success is inferred from complete files.

Memory and CPU. Sampled device memory is the maximum NVML used bytes during the block. Parent-process allocator peaks are recorded separately. They are zero for the vLLM paths and are treated there as unavailable. Own-process CPU time is sampled over marked windows that include warmup and idle intervals, with coverage recorded. It is reported as sampled cost, not as isolated inference CPU time or CPU energy.

Foreign CPU load. Host covariates are sampled every second during active visits, after model and worker setup. Node busy and total ticks come from interval deltas of the kernel CPU statistics over the 96 logical CPUs of NUMA node 1 (48–95 and 144–191). Benchmark-own ticks are the user plus system time, over the same intervals, of the measured process tree and registered runtime workers. Each process is identified by process identifier and start time. The foreign fraction is node busy ticks minus benchmark-own ticks, divided by node total ticks. Total, own and raw residual series are all retained, and only negligible negative round-off is clamped for presentation. Benchmark-own CPU time is measured cost, not contamination. Incomplete accounting, inconsistent counters or an own worker running outside NUMA node 1 make the foreign covariate unavailable for that interval, not zero.

Cross-visit interpretation. A contrast between two visits is interpretable when four conditions hold. Their median foreign NUMA-node CPU fractions differ by at most 10 percentage points. Their median GPU 3 utilizations differ by at most 5 percentage points. The sets of processes observed on GPU 3 match. Each visit has valid covariates and at least two clean blocks. Missing covariates make the contrast uninterpretable, and the time gap between visits is reported. For an interpretable contrast, a direction is stated only when the absolute difference of block medians exceeds the larger of the two visits' within-visit block ranges (maximum minus minimum). Otherwise the values are reported without an ordering. The within-visit range cannot estimate or bound between-visit variation from model reloads, allocator or cache state, clocks or thermal drift. The load covariates are operational filters and do not prove an uncontended host.

Sessions and bridge. The bridge cell is BANKING77 known-label constrained decoding at batch size 32 with certification inside timed calls. It was measured once in each session on the same verified GPU, and its repaired-session measurement is the before visit of the optimization trial. Its median throughputs are 25.84 and 25.85 decisions/s, and its larger within-visit range is 0.334 decisions/s. Consider a contrast between an original-session and a repaired-session visit. A direction for a given metric is retained only when the contrast's difference exceeds the bridge's between-session shift in that metric. When the bridge lacks the metric, the direction is suppressed. The screen can only suppress directions. The artifact lists each of the 100 registered contrasts with its screen outcome.

I SHARED-STATE EXPERIMENT

The pinned, MIT-licensed fixture (TheoLeeCJ, 2026) contains 37 authored states with 21 questions each and no gold labels. The complete 21-question fixture defines the question order and the state ranking. States are sorted by UTF-8 state byte length and then group identifier, and zero-based ranks 0 (8,004 bytes) and 18 (8,103 bytes) are measured. The $Q = 1$ and $Q = 16$ bundles take the first one and sixteen questions of each selected state. They are verified to equal the corresponding prefixes of the full question order, and no 21-question timing cell is run.

Each decision is read from next-token probabilities over the declared single-token candidate identifiers. All paths use scaled dot-product attention in eager execution, and no compilation benefit is assumed. The fresh paths make direct serial or batched decisions. The reuse paths run one state-prefix prefill followed by serial or parallel question branches. The prefix is tokenized so that prefix and branch tokens reproduce the complete-prompt token sequence, and its KV state is copied or materialized for the branches. Caches are discarded after each bundle, so zero-copy KV sharing and persistent cross-request caching are outside the scope. No path generates a serialized probability string.

Numerical checks on the GPU runtime compare normalized candidate log probabilities and selected identifiers across the four paths. They use four input views of 37, 148, 592 and 777 requests, grouped into complete same-state bundles of 1, 4, 16 and 21 questions. FP32 uses $\mathrm{atol} = 10^{-4}$

and $\mathrm{rtol} = 10^{-5}$, and all 12 FP32 same-precision comparisons meet these tolerances over 4,662 dependent request comparisons. BF16 autocast paths were compared with $\mathrm{atol} = 0.1$ and $\mathrm{rtol} = 0.002$, both with each other and directly with FP32, because tolerance passes are not transitive. Only the $Q = 1$ fresh batched comparison passes among the 12 BF16 same-precision comparisons, and none of the 16 BF16-versus-FP32 comparisons passes. These comparisons remain numerical evidence only and support no equivalence or reuse-benefit claim. CPU checks on small models and structural configuration checks are preparatory. They provide no evidence of numerical agreement or speed for the pretrained model.

J ONE-FACTOR CELLS

(state tokens, K, B)	$(512, 8, 1)$	$(2048, 8, 1)$	$(512, 2, 1)$	$(512, 32, 1)$	$(512, 8, 8)$	$(512, 8, 32)$
Adapted constrained	164.0	66.2	169.9	141.5	268.0	316.9
Released likelihood†	22.5	5.2	82.5	5.6	24.8	30.3
Released cached likelihood	71.0	33.1	94.9	30.6	122.4	140.1
Query head	46.8	38.1	47.3	46.2	268.1	439.4
Set head	46.5	38.0	47.7	45.3	266.7	446.7

Table 8: Median throughput (decisions/s over five blocks) of the 30 one-factor cells on constructed BANKING77-derived 32-request cohorts. Each column changes one factor from the anchor $(512, 8, 1)$. All cells of one method come from the same visit. Released decoders are Qwen3-0.6B, and the heads are fully adapted ModernBERT-base at seed 17. †Cells from the visit classified as incomplete execution. Block ranges, latencies, energy and memory are in the artifact. The values are descriptive cost evidence without a semantic accuracy claim or interaction estimate.

K REPRODUCIBILITY DETAILS

Table 9 lists pinned public model revisions. Training seeds are 17, 29 and 43. Seed 1701 is used for candidate sampling in training, the dynamic-cohort hash, candidate permutations and the runtime visit order. The changing-candidate group selection uses the salted digest of Appendix E. Hardware, device placement, CPU affinity and stopping rules are given in Section A.2.5 and Appendix H.

Model	Revision
Qwen/Qwen3-0.6B	`c1899de289a04d12100db370d81485cdf75e47ca`
Qwen/Qwen3-1.7B	`70d244cc86ccca08cf5af4e1e306ecf908b1ad5e`
Qwen/Qwen3-4B	`1cfa9a7208912126459214e8b04321603b3df60c`
answerdotai/ModernBERT-base	`8949b909ec900327062f0ebf497f51aef5e6f0c8`
answerdotai/ModernBERT-large	`45bb4654a4d5aaff24dd11d4781fa46d39bf8c13`
Knowledgator/gliclass-base-v3.0	`77a70e6cd52e602ed18184ef37d18bdd3741e3d5`
Knowledgator/gliclass-modern-base-v3.0	`ac369222ca4375ca66ebaf7fb5220f223514c035`

Table 9: Pinned public model revisions used by the compared endpoints.